Efficient designs for mean estimation in multilevel populations and test norming

Chapter 7

Summary

This thesis deals with sample size planning and optimal design for two types of observational studies: (1) surveys for mean estimation in multilevel populations, such as school-based surveys for estimating mean alcohol consumption among high school students (see, for instance, ESPAD Group, 2016), and (2) normative studies to derive reference values for tests and questionnaires, such as neuropsychological tests to assess information processing speed (see, for instance, Van der Elst et al., 2006a), and clinical questionnaires to measure patients' orientation toward chronic pain (see, for instance, Van Breukelen & Vlaeyen, 2005).

Chapter 1 provides an introduction to these two types of studies. Specifically, the practical importance of these studies is explained, real-life examples are provided, the main results available in the literature are summarized, and the statistical models for analysing data obtained with these studies are introduced. Furthermore, a definition of optimal design for each type of study is given, and strategies to find robust designs are presented. Chapter 1 ends with an outline of the thesis.

Chapter 2 is on unbiased and efficient estimation of the average of all individual outcomes in two-level populations, with either simple random sampling (SRS) of individuals (i.e. individuals are drawn directly from the population) or two-stage sampling (TSS) (i.e. first clusters are sampled, and then individuals are sampled from the selected clusters). Cluster sizes are allowed to vary and to be related to the outcome variable of interest (i.e. cluster size is informative). Three TSS schemes are considered: sampling clusters with probability proportional to cluster size and then taking a SRS of the same number of individuals within each sampled cluster (TSS1); drawing a SRS of clusters and then sampling the same percentage of individuals per cluster (TSS2); taking a SRS of clusters and then the same number of individuals per sampled clusters (TSS3). In this chapter, it is shown that the average of all

individual outcomes and the average of all cluster-specific means (i.e. the two definitions of population means in a two-level population) coincide only if cluster sizes are either equal or non-informative. Unbiased estimation of the average of all individual outcomes is discussed under each sampling scheme. Furthermore, the three TSS schemes are compared in terms of efficiency with each other and with SRS of individuals, under the constraint of a fixed total sample size. The relative efficiency of the sampling schemes is shown to vary across different cluster size distributions. However, TSS1 is the most efficient TSS scheme for many cluster size distributions. Model-based and design-based inference are compared and are shown to give similar results, at least if the model assumptions are met. The results of this chapter are applied to two real-life cluster size distributions, that is, the distribution of high school size in Italy, and the distribution of patient list size in England.

Chapter 3 deals with optimal TSS schemes for mean estimation in two-level populations, when cluster size is informative. A simulation study is performed to assess the bias in the sampling variance formulas derived for TSS2 and TSS3 in chapter 2, because these variances are based on approximations. Optimal sample size equations are derived for each TSS scheme considered in chapter 2. The optimal design is the number of clusters and number of individuals per cluster that minimizes the sampling variance of the population mean estimator, subject to a cost constraint. The consequences for the optimal design of ignoring informative cluster size are investigated. It turns out that the optimal TSS designs are quite robust against misspecification of the degree of informativeness of cluster size, but assuming non-informative cluster size can lead to serious underestimation of the required research budget for the desired power level. Furthermore, the three optimal TSS schemes are compared, in terms of efficiency, with each other and with SRS of individuals, under the constraint of a fixed budget for sampling and measuring. The optimal TSS1 is shown to be the most efficient sampling scheme for many cluster size distributions. To overcome the dependency of the optimal designs on the prior knowledge of some model parameters, maximin designs are derived for each TSS scheme. Finally, a procedure is proposed to derive maximin sample sizes and a maximin budget split between two surveys to estimate and compare the means of two populations with TSS1. This procedure is illustrated when planning a hypothetical survey to compare adolescent alcohol consumption between France and Italy, using the real distributions of high school size in these two countries.

Chapter 4 focuses on normative studies to derive reference values or norms for tests and questionnaires. In this chapter, the regression-based approach to norming is adopted, because it

has three advantages over the traditional approach of norming per subgroup defined in terms of age and sex (or other demographic variables): (i) it is more efficient, that is, it requires a smaller sample size, (ii) it allows to identify the predictors that affect the test score of interest, and (iii) it allows to derive the optimal design, which is defined as the joint distribution of scores on the predictors that minimizes the sampling variance of the norm statistic under the assumed norming model. In chapter 4, sampling variance formulas are derived for commonly used norm statistics, that is, Z-scores and percentile rank scores. Since these variance formulas are based on approximations, two simulation studies are performed to assess the bias induced by these approximations. These sampling variance formulas are used to derive optimal designs for five regression models including a quantitative and a qualitative predictor, differing in whether they allow for interaction and nonlinearity. Efficient designs that are robust against misspecification of the norming model are derived using the maximin strategy with efficiency and relative efficiency as criteria. It is shown that, for the five considered regression models, the most robust designs obtained under these two criteria are the same. Furthermore, for the optimal design formulas are proposed to determine the required size of the normative sample for each norm statistic (i.e. Z-score and percentile rank score). These formulas can be used to ensure either the desired power level for hypothesis testing or the desired margin of error for interval estimation. The results of this chapter are illustrated using Van der Elst et al. (2006b)'s normative study of the Profession Naming verbal fluency test.

Chapter 5 extends chapter 4 to a scenario of several tests to be normed with the same sample. To take into account the correlation between test scores of the same individual, a multivariate regression model is used (Van der Elst et al., 2017), instead of estimating a univariate regression model for each test. However, in the multivariate regression-based approach of Van der Elst et al. (2017), each test is normed separately, thus ignoring the correlation between norm statistic values of the same individual. In chapter 5, a new multivariate regression-based approach is proposed that combines all separate scores for an individual in the Mahalanobis distance (i.e. between the multivariate test score for an individual and the multivariate average in the reference population), thus providing an indicator of the individual's overall performance across all tests. Furthermore, sampling variance and covariance formulas are derived for the Z-score estimator, as well as a sampling variance formula for the Mahalanobis distance estimator. Since all these formulas are based on approximations, two simulation studies are performed to assess the bias induced by these approximations, thus extending the results of the simulation studies in chapter 4 to the case of

two tests to be normed. For both multivariate regression-based approaches, optimal designs are derived for the multivariate version of the five regression models as considered in chapter 4, and efficient designs that are robust against model misspecification are obtained using the maximin strategy with efficiency and relative efficiency as criteria. It is shown that the most robust designs obtained under the two criteria coincide. Formulas to derive the required size for the optimal design of the normative sample are proposed for the Mahalanobis distance-based approach only, because Van der Elst et al. (2017)'s approach is hampered by multiple testing issues. The results of this chapter are illustrated using Van der Elst et al. (2006a)'s normative study of the oral and written versions of the Letter Digit Substitution Test.

Chapter 6 has provided some considerations about common challenges that are encountered in planning surveys and normative studies. Furthermore, a few practical guidelines on how to design each type of study have been given, and ideas for future research have been discussed.

In the next chapter, a reflection is given on the scientific and social impact of this thesis.

Samenvatting

Efficiënte ontwerpen voor gemiddelde schatting in multilevel populaties en testnormering

Dit proefschrift behandelt de planning van de steekproefgrootte en het optimale ontwerp (design) voor twee soorten observationele studies: (1) surveys voor het schatten van een gemiddelde in multilevel populaties, zoals de gemiddelde alcoholconsumptie onder middelbare scholieren (zie, bijvoorbeeld, ESPAD Group, 2016) en (2) normatieve studies om referentiewaarden af te leiden voor tests en vragenlijsten, zoals neuropsychologische tests om de snelheid van de informatieverwerking te beoordelen (zie, bijvoorbeeld, Van der Elst et al., 2006a), en klinische vragenlijsten om de oriëntatie van patiënten op chronische pijn te meten (zie, bijvoorbeeld, Van Breukelen & Vlaeyen, 2005).

Hoofdstuk 1 geeft een inleiding tot deze twee soorten studies. Het praktische belang van deze studies wordt uitgelegd, er worden praktijkvoorbeelden gegeven, de belangrijkste resultaten in de literatuur worden samengevat en de statistische modellen voor het analyseren van gegevens die met deze studies zijn verkregen worden geïntroduceerd. Verder wordt een definitie gegeven van het optimale ontwerp (optimal design) voor elk soort studie en worden strategieën gepresenteerd om robuuste ontwerpen te vinden. Hoofdstuk 1 eindigt met een overzicht van het proefschrift.

Hoofdstuk 2 gaat over zuivere en efficiënte schatting van het gemiddelde van alle individuele uitkomsten in twee-niveau populaties, met ofwel simple random sampling (SRS) van individuen (d.w.z. individuen worden rechtstreeks uit de populatie getrokken) of two-stage sampling (TSS) (d.w.z. eerst worden clusters getrokken, en vervolgens worden individuen uit de geselecteerde clusters getrokken). Clustergroottes mogen variëren en gerelateerd zijn aan de uitkomstvariabele van belang (d.w.z. de clustergrootte is informatief). Er worden drie TSS methoden onderzocht: Het trekken van clusters met een waarschijnlijkheid evenredig aan de clustergrootte, en het vervolgens trekken van een SRS van hetzelfde aantal individuen binnen elk getrokken cluster (TSS1); Het trekken van een SRS van clusters en het vervolgens trekken van hetzelfde percentage individuen per getrokken cluster (TSS2); Het trekken van een SRS

van clusters en het vervolgens trekken van hetzelfde aantal individuen per getrokken cluster (TSS3). In dit hoofdstuk wordt aangetoond dat het gemiddelde van alle individuele uitkomsten en het gemiddelde van alle clusterspecifieke gemiddelden (d.w.z. de twee definities van populatiegemiddeldes in een twee-niveau populatie) alleen samenvallen als de clustergrootten gelijk of niet-informatief zijn. Zuivere schatting van het gemiddelde van alle individuele uitkomsten wordt voor elke steekproefmethode (sampling scheme) besproken. Bovendien worden de drie TSS-methoden in termen van efficiëntie met elkaar en met SRS van individuen vergeleken, uitgaande van een vastgestelde totale steekproefomvang. Aangetoond wordt dat de relatieve efficiëntie van de sampling schemes afhangt van de verdeling van de clustergrootte. TSS1 is echter de meest efficiënte TSS methode voor veel clustergrootte verdelingen. Model-based en design-based inference worden met elkaar vergeleken en er wordt aangetoond dat zij vergelijkbare resultaten opleveren indien aan de modelaannames wordt voldaan. De resultaten van dit hoofdstuk worden toegepast op twee echte clustergrootte verdelingen: de verdeling van de middelbare schoolgrootte in Italië, en de verdeling van de grootte van huisartspraktijken in Engeland.

Hoofdstuk 3 gaat over optimale TSS designs voor het schatten van het gemiddelde in twee-niveau populaties wanneer de clustergrootte informatief is. Er wordt een simulatiestudie uitgevoerd om de bias te beoordelen in de steekproefvariantie (sampling variance) formules die zijn afgeleid voor TSS2 en TSS3 in hoofdstuk 2, omdat deze varianties gebaseerd zijn op benaderingen. Voor elke TSS methode die in hoofdstuk 2 wordt beschouwd, worden vergelijkingen voor de optimale steekproefomvang afgeleid. Het optimale ontwerp (optimal design) is het aantal clusters en het aantal personen per cluster dat de steekproefvariantie (sampling variance) van de schatter van het populatiegemiddelde minimaliseert, onder een kostenbeperking. De gevolgen voor het optimale ontwerp (optimal design) van het negeren van informatieve clustergrootte worden onderzocht. Het blijkt dat de optimale TSS-ontwerpen heel robuust zijn tegen misspecificatie van de mate van informativiteit van clustergrootte, maar aannemen dat clustergrootte niet informatief is kan leiden tot ernstige onderschatting van het vereiste onderzoeksbudget voor het gewenste power niveau. Bovendien worden de drie optimale TSS methoden in termen van efficiëntie met elkaar en met SRS van individuen vergeleken uitgaande van een vastgesteld budget voor steekproeftrekking en meting. Aangetoond wordt dat de optimale TSS1 het meest efficiënte design is voor veel clustergrootte verdelingen. Om de afhankelijkheid van de optimale ontwerpen (optimal designs) van voorkennis over sommige modelparameters te overwinnen, worden maximin ontwerpen

(maximin designs) afgeleid voor elke TSS methode. Ten slotte wordt een procedure voorgesteld om maximin steekproefomvangen van, en maximin budgetverdeling (budget split) tussen, twee surveys af te leiden om de gemiddelden van twee populaties te schatten en te vergelijken met TSS1. Deze procedure wordt geïllustreerd met de planning van een hypothetische survey om het alcoholgebruik van adolescenten te vergelijken tussen Frankrijk en Italië, gebruikmakend van de echte verdelingen van de middelbare schoolgrootte in deze twee landen.

Hoofdstuk 4 richt zich op normatieve studies om referentiewaarden of normen voor tests en vragenlijsten af te leiden. In dit hoofdstuk wordt de regression-based benadering van normering gevolgd, omdat die drie voordelen heeft ten opzichte van de traditionele benadering van normering per subgroep gedefinieerd in termen van leeftijd en geslacht (of andere demografische variabelen): (i) de regression-based benadering is efficiënter, dat wil zeggen, deze vereist een kleinere steekproefomvang, (ii) en maakt het mogelijk om de voorspellers te identificeren die de testscore beïnvloeden, en (iii) en maakt het mogelijk om het optimale ontwerp (optimal design) af te leiden, dat wordt gedefinieerd als de gezamenlijke verdeling van scores op de voorspellers die de steekproefvariantie (sampling variance) van de referentiewaarde of norm onder het veronderstelde regressiemodel minimaliseert. In hoofdstuk 4 worden steekproefvariantie (sampling variance) formules afgeleid voor veelgebruikte normen nl. Z-scores en percentile rank scores. Aangezien deze variantieformules gebaseerd zijn op benaderingen, worden twee simulatiestudies uitgevoerd om de bias te beoordelen die door deze benaderingen wordt veroorzaakt. Deze steekproefvariantie (sampling variance) formules worden gebruikt om optimale ontwerpen (optimal designs) af te leiden voor vijf regressiemodellen met daarin een kwantitatieve en kwalitatieve voorspeller, die verschillen in of ze interactie en niet-lineariteit toelaten. Efficiënte ontwerpen (efficient designs) die robuust zijn tegen misspecificaties van het regressiemodel worden afgeleid met behulp van de maximin strategie (maximin strategy), met efficiëntie en relatieve efficiëntie als criteria. Aangetoond wordt dat voor de vijf onderzochte regressiemodellen de meest robuuste ontwerpen (robust designs) die voor deze twee criteria verkregen zijn, dezelfde zijn. Bovendien worden voor het optimale ontwerp (optimal design) formules voorgesteld om de vereiste omvang van de normatieve steekproef voor elke norm (d.w.z. Z-score en percentile rank score) te bepalen. Deze formules kunnen worden gebruikt om het gewenste power niveau voor hypothese toetsing of de gewenste foutmarge (margin of error) voor intervalschatting te garanderen. De resultaten van dit hoofdstuk worden geïllustreerd aan de hand van een normatieve studie van de Profession Naming Verbal Fluency Test (zie Van der Elst et al., 2006b).

Hoofdstuk 5 breidt hoofdstuk 4 uit tot een scenario van verschillende tests die met dezelfde steekproef moeten worden genormeerd. Om rekening te houden met de correlatie tussen de testscores van hetzelfde individu, wordt een multivariaat regressiemodel gebruikt (Van der Elst et al., 2017), in plaats van een univariaat regressiemodel voor elke test te schatten. In de multivariate regression-based benadering van Van der Elst et al. (2017) wordt elke test echter afzonderlijk genormeerd, waardoor de correlatie tussen normwaarden van hetzelfde individu wordt genegeerd. In hoofdstuk 5 wordt een nieuwe multivariate regression-based benadering voorgesteld die alle afzonderlijke scores voor een individu in de Mahalanobis-distance (d.w.z. de afstand tussen de multivariate testscore voor een individu en het multivariate gemiddelde in de referentiepopulatie) combineert, waardoor een indicator van de algehele prestatie van het individu op alle tests wordt verkregen. Bovendien worden steekproefvariantie (sampling variance) en covariantie formules afgeleid voor de Z-score schatter, evenals een steekproefvariantie (sampling variance) formule voor de Mahalanobis-distance schatter. Aangezien al deze formules gebaseerd zijn op benaderingen, worden twee simulatiestudies uitgevoerd om de door deze benaderingen veroorzaakte bias te beoordelen, waarbij de resultaten van de simulatiestudies in hoofdstuk 4 worden uitgebreid tot het geval van twee te normeren tests. Voor beide multivariate regression-based benaderingen worden optimale ontwerpen (optimal designs) afgeleid voor de multivariate versie van de vijf regressiemodellen zoals beschouwd in hoofdstuk 4, en worden efficiënte ontwerpen (efficient designs) die robuust zijn tegen modelmisspecificatie verkregen met behulp van de maximin strategie (maximin strategy), met efficiëntie en relatieve efficiëntie als criteria. Aangetoond wordt dat de meest robuuste ontwerpen (robust designs) die op grond van de twee criteria zijn verkregen, samenvallen. Formules om de vereiste omvang af te leiden voor het optimale ontwerp (optimal design) van de normatieve steekproef worden alleen gepresenteerd voor de Mahalanobis distance-based benadering, omdat de benadering van Van der Elst et al. (2017) gehinderd wordt door het probleem van multiple testing. De resultaten van dit hoofdstuk worden geïllustreerd aan de hand van de normatieve studie van de mondelinge en schriftelijke versies van de Letter Digit Substitution Test (zie Van der Elst et al., 2006a).

Hoofdstuk 6 geeft een aantal overwegingen over gemeenschappelijke uitdagingen die worden ondervonden bij het ontwerpen van enquêtes (surveys) en normatieve studies. Verder worden er enkele praktische richtlijnen gegeven voor het ontwerpen van elk soort studie en worden ideeën voor toekomstig onderzoek besproken.